8 APPENDIX

In the following, we will prove that it is difficult and sometimes even impossible to set a proper confidence threshold to make the distinguish between $\mathbf{M}_{all}$ and $\mathbf{M}_{mix}$.

First, based on the definitions, if R_{fp} and R_{fn} satisfy the uniform-random distribution, the confidence of root-cause nodes $\mathbf{M}_{all}$ and their children nodes can be computed as

$$\mathcal{C}_{all} = \frac{Leaf_{all} \times (1 - R_{fn})}{Leaf_{all}} = 1 - R_{fn} \tag{1}$$

The confidence of the first kind of non-root-cause nodes $\mathbf{M}_{none}$ can be computed as

$$\mathcal{C}_{none} = \frac{Leaf_{none} \times R_{fp}}{Leaf_{none}} = R_{fp} \tag{2}$$

For the second kind of non-root-cause nodes $\mathbf{M}_{mix}$, the γ percentage of leaf nodes descended from $\mathbf{M}_{mix}$ are descended from the root-cause nodes $\mathbf{M}_{all}$, while $1 - \gamma$ percentage of leaves descended from $\mathbf{M}_{mix}$ are also descended from the non-root-cause nodes $\mathbf{M}_{none}$. For example, if $\mathcal{M}_1 = \{d1, l1, *\}$ is the unique root-cause node $\mathbf{M}_{all}$, $\mathcal{M}_2 = \{d2, l1, *\}$ is a non-root-cause node $\mathbf{M}_{none}$ and $\mathcal{M}_3 = \{*, l1, *\}$ is a non-root-cause node $\mathbf{M}_{mix}$. Then the confidence of $\mathbf{M}_{mix}$ can be computed as

$$\begin{aligned}\mathcal{C}_{mix} &= \frac{Leaf_{mix} \times \gamma \times (1 - R_{fn}) + Leaf_{mix} \times (1 - \gamma) \times R_{fp}}{Leaf_{mix}} \\ &= \gamma - \gamma \times R_{fn} + R_{fp} - \gamma \times R_{fp}\end{aligned}$$

We use a confidence threshold δ_2 to filter the graph nodes with low *confidence* value. If an ideal anomaly detection with 100% accuracy is applied, the *confidence* is 100% for both the root-cause nodes $\mathbf{M}_{all}$ and their subset, while the *confidence* is 0% for all other non-root-cause nodes $\mathbf{M}_{none}$. If $\delta_2 > (1 - R_{fn})$, then we will miss all the root-cause nodes $\mathbf{M}_{all}$. If $\delta_2 < R_{fp}$, then we will report the non-root-cause nodes $\mathbf{M}_{none}$ as root-cause nodes falsely. Then an effect confidence threshold δ_2 should satisfy $R_{fp} < \delta_2 < 1 - R_{fn}$ and $R_{fp} + R_{fn} < 1$.

It is difficult to find an appropriate confidence threshold δ_2 to identify the non-root-cause nodes $\mathbf{M}_{mix}$ and their children nodes correctly. It can be proved that if $R_{fp} + R_{fn} < 1$ then $\mathcal{C}_{none} < \mathcal{C}_{mix} < \mathcal{C}_{all}$ i.e., $R_{fp} < \mathcal{C}_{mix} < 1 - R_{fn}$. We first prove that $\mathcal{C}_{mix} < \mathcal{C}_{all}$ as follows:

$$\begin{gathered}\because R_{fp} + R_{fn} < 1 \Longleftrightarrow R_{fp} < 1 - R_{fn} \\ \because 1 - \gamma > 0 \\ \therefore (1 - \gamma) \times R_{fp} < (1 - \gamma) \times (1 - R_{fn}) \\ \Longleftrightarrow \mathcal{C}_{mix} < \mathcal{C}_{all}\end{gathered} \tag{3}$$

Similarly,

$$\begin{gathered}\because R_{fp} + R_{fn} < 1 \Longleftrightarrow 1 - R_{fn} > R_{fp} \\ \because \gamma > 0 \\ \therefore \gamma \times (1 - R_{fn}) > \gamma \times R_{fp} \\ \therefore \gamma \times (1 - R_{fn}) + (1 - \gamma) \times R_{fp} > R_{fp} \\ \Longleftrightarrow \mathcal{C}_{mix} > \mathcal{C}_{none}\end{gathered} \tag{4}$$

$$r^1 = \sqrt{(x^1 - x)^2 + (y^1 - y)^2 + (z^1 - z)} \tag{5}$$

$$r^2 = \sqrt{(x^2 - x)^2 + (y^2 - y)^2 + (z^2 - z)} \tag{6}$$

$$r^3 = \sqrt{(x^3 - x)^2 + (y^3 - y)^2 + (z^3 - z)} \tag{7}$$

With γ changing between $(0, 1)$ constantly, $\mathcal{C}_{mix}$ changes with $(R_{fp}\ , 1 - R_{fn})$ constantly. Since the R_{fp} and R_{fn} can not be a perfect uniform-random distribution in practice, there is no clear boundary between $\mathbf{M}_{all}$ and $\mathbf{M}_{mix}$. This completes the proof. Hence it is difficult to judge whether a non-root-cause node $\mathbf{M}_{mix}$ is a root-cause node or not based on the confidence metric.

Therefore, it is impossible to simply use $1 - R_{fn}$ as λ_c to distinguish between $\mathbf{M}_{all}$ and $\mathbf{M}_{mix}$. We have confidence threshold λ_c which satisfies $R_{fp} < \lambda_c < 1 - R_{fn}$. Then we can easily remove $\mathbf{M}_{none}$ whose confidence metric is smaller than λ_c from the candidates. All the nodes $\mathbf{M}_{all}$ and some nodes $\mathbf{M}_{mix}$ whose confidence metric is larger than λ_c will form a candidate set.